Goal 5

Carter Rogers, Nicholas Hunter, Shivam K C

10/25/2019

# Our Problem and Hypothesis:

Some countries are happier than others. We want to find what factors best predict happiness. Therefore, we hypothesize the following: (i) We expect factors such as life expectancy and GDP to be positively correlated with happiness. In our dataset, Life Ladder is the happiness score of a nation gotten by averaging the national response to survey, Cantril Ladder, to a single number. (ii) We expect factors such as corruption to be negatively correlated with happiness. (iii) We expect Europe to be the happiest continent on average.

#Loading Libraries  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

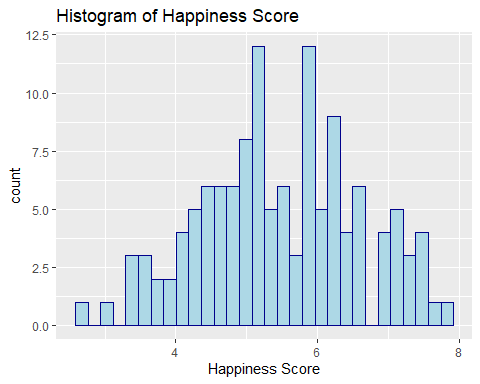
library(reshape2)

# Import main dataset   
WHR = read.csv("World\_Happiness\_Report\_2008-2018.csv")  
  
# Filter most recent year  
WHR2018 = WHR %>% filter(Year == 2018)  
  
#Changing column names  
names(WHR2018)[names(WHR2018) == "ï..Country.name"] <- "Country.name"  
  
# Load continents  
# NA == North America so need to specify na.strings  
countries\_and\_continents = read.csv("countries and continents.csv", na.strings="N/A")  
  
  
# Merge main dataset with continent data  
whr2018\_w\_continents = merge(x=WHR2018, y=countries\_and\_continents, by.x='Country.name', by.y='name')

# Distribution of Happiness Score

ggplot(data=whr2018\_w\_continents, aes(x=Life.Ladder))+geom\_histogram(  
 color="darkblue", fill="lightblue")+ggtitle("Histogram of Happiness Score")+xlab("Happiness Score") #histogram code

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



We plotted this graph to observe the distribution of happiness score. Here, the distribution of happiness score is symmetric. This means that there are many countries with average happiness scores.

# Statistics of Happiness Score

summary(whr2018\_w\_continents$Life.Ladder) #summary of happiness score

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.694 4.661 5.409 5.469 6.263 7.858

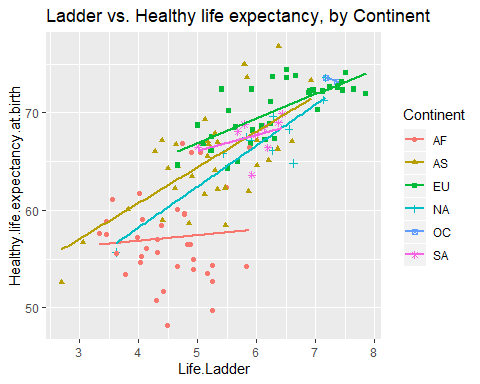
We are using this table to check the statistics (spread and mean) of Happiness Score which is our dependent variable. We observe that the mean Happiness Score of the World in 2018 was 5.469. The maximum score was 7.86 and the minimum score was 2.69.

# Testing Hypothesis 1

ggplot(whr2018\_w\_continents, aes(x=Life.Ladder, y=Healthy.life.expectancy.at.birth, shape=Continent, color=Continent))+geom\_point()+ggtitle("Ladder vs. Healthy life expectancy, by Continent")+geom\_smooth(method=lm, se=FALSE) #Relationship between happiness score and life expectancy among different continents.

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).

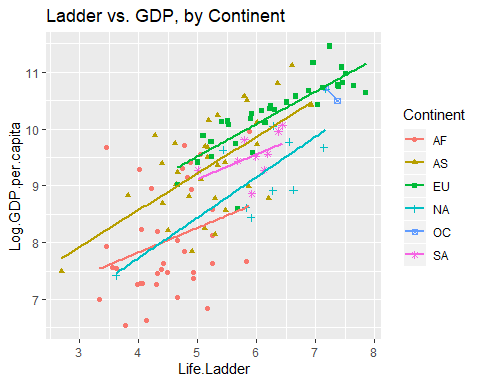


We constructed this plot in order to examine the relationship between healthy life expectancy for each country and that respective country’s happiness score. We see that European and Asian countries are strongly and positively correlated with happiness. African countries are not as strongly correlated, as the line of best fit has a lower slope than the others.

ggplot(whr2018\_w\_continents, aes(x=Life.Ladder, y=Log.GDP.per.capita, shape=Continent, color=Continent))+geom\_point()+ggtitle("Ladder vs. GDP, by Continent")+geom\_smooth(method=lm, se=FALSE) #Relationship between happiness score and GDP among different continents.

## Warning: Removed 5 rows containing non-finite values (stat\_smooth).

## Warning: Removed 5 rows containing missing values (geom\_point).



Similar to the previous plot, we constructed this plot to examine the relationship between GDP and happiness score. We can see that for all continents, GDP is strongly and positively correlated with happiness. In particular, European countries typically have both higher GDP’s and happiness scores.

# Testing Hypothesis 3

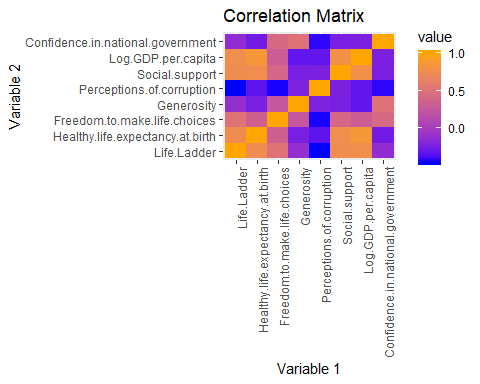
happiness\_by\_continent <- whr2018\_w\_continents %>% group\_by(Continent) %>% summarise(count = n(), ladder\_avg = mean(Life.Ladder)) #piping to get average happiness scores for each continent.  
happiness\_by\_continent

## # A tibble: 6 x 3  
## Continent count ladder\_avg  
## <fct> <int> <dbl>  
## 1 AF 37 4.52  
## 2 AS 34 5.17  
## 3 EU 35 6.36  
## 4 NA 10 6.08  
## 5 OC 2 7.27  
## 6 SA 9 5.94

We made this table to test our hypothesis 3. We observe Oceanic to have the highest Happiness score, but this score only considers 2 countries. We do observe (excluding OC) Europe to be the happiest continent.

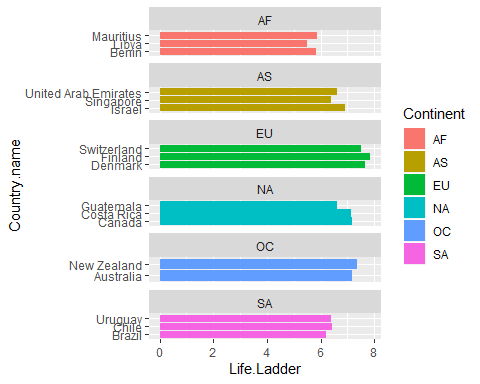
# Testing Hypothesis 2

# Select columns of interest  
whr\_2018\_select = whr2018\_w\_continents %>% select(Life.Ladder, Healthy.life.expectancy.at.birth, Freedom.to.make.life.choices, Generosity, Perceptions.of.corruption, Generosity, Social.support, Log.GDP.per.capita, Confidence.in.national.government)  
  
# Create correlation matrix  
cormat <- round(cor(whr\_2018\_select, use = "complete.obs"),2)  
  
# Melt the matrix  
melted\_cormat = melt(cormat)  
  
# Plot the correlation data  
ggplot(data = melted\_cormat, aes(x=Var1, y=Var2, fill=value)) + xlab("Variable 1") + ylab("Variable 2") + ggtitle("Correlation Matrix") + geom\_tile()+theme(axis.text.x=element\_text(angle=90, hjust=1)) + scale\_fill\_gradient(low = "blue", high = "orange")

 This plot shows the correlations between a selected set of variables which we believe could relate to our hypothesis. Here we see stronger correlations between Life.Ladder, which is our self reported happiness score, and GDP per capita, Social Support, Freedom, and Life expectancy. We see weaker correlations with Confidence in Government and Generosity. We also see a negative correlation between happiness and Perceptions of Corruption.

# Comparing Top 3 Happiest Countries from Each Continent

Top\_3 <- whr2018\_w\_continents %>% select(Continent, Country.name,Life.Ladder) %>% group\_by(Continent)%>% arrange(desc(Life.Ladder), .by\_group = TRUE)%>% slice(1:3) #getting the table  
  
#Now plotting bar plots and comparing  
  
ggplot(Top\_3,aes(x=Country.name, y=Life.Ladder, fill = Continent))+geom\_bar(stat='identity')+coord\_flip()+facet\_wrap(~Continent,ncol=1,scales = 'free\_y')



We plotted this graph to compare top 3 happiest countries from each continent. We observe that top 3 happiest countries from Europe are happier than or at the (very least) same level as top 3 happiest countries from other continents.